CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Collection

10-12 October 2017, Ottawa, Canada

WP.1-3

16 August 2017

Surveying Informal Enterprises: Applying Stratified Adaptive Cluster Sampling using CAPI with Implementation and Monitoring Tools

Gemechu Aga , David C Francis and Michael Wild (World Bank)

gayanaaga@worldbank.org, dfrancis@worlbank.org and mwild@worldbank.org

Abstract

Hard-to-reach populations (e.g., informal sector establishments, pastoralist households) are difficult to survey since they are missing from standard sampling frames and their location is usually not fixed. Additionally, they are often geographically clustered. This paper discusses a sampling approach that addresses these two key challenges. We use a stratified (spatial) adaptive cluster approach for a pilot survey on informal firms in Harare, Zimbabwe. As in standard area based sampling our Primary Sampling Units are geographic clusters. In our case, we create a spatial sampling frame by partitioning the city of Harare into a grid of 200 by 200 meter cells. This enables probability surveys even in the absence of a comprehensive and current list of firms. To account for clustering of informal firms, selection of grid cells) is adaptive. A starting square (grid cell) is randomly selected and all informal firms in the square are enumerated. If the number of informal firms in the starting square exceeds a predefined threshold, all other squares surrounding the starting square are surveyed. This approach has rarely been applied in face-to-face establishment surveys due to challenges in field management. However, with the World Bank's Survey Solution software package and the use of navigation software and real-time field monitoring, we successfully implemented this sampling design for the said pilot survey of informal establishments.

1. Introduction

A requirement for a standard probability survey is the existence of a comprehensive sampling frame, which allows for the calculation of weights and, subsequently, for the calculation of population estimates. However, for many populations of interest, an adequate and comprehensive frame does not exist. Populations like informal sector units, pastoralist households, or homeless people are therefore often not included in standard surveys.

Nevertheless, these populations constitute a significant share of the overall population in some countries, and having them excluded from the statistical system precludes the possibility of well targeted public policies as well as the measurement of any progress in their living conditions. Specifically, the informal sector is known to be a predominant source of employment and income in many developing countries (De Soto 1989), yet these jobs are generally held as less desirable, more unstable, and lower paying (Harris and Todaro 1970; La Porta and Schleifer 2014). Simultaneously, understanding the operation, activity, and productivity of units (or firms) in this sector is a key focus for researchers and policymakers alike. As such, the measurement of informal sector activity also directly relates to the correct measurement of the World Bank Group's twin goals of reducing extreme poverty and decreasing income inequality, because only by having good quality statistics, are policymakers in the position to determine any progress or achievement.

Non-coverage of some segments of the population by a sampling frame can seriously affect the production of correct population estimates, and in the worst-case scenario — if the non-covered sub-population is significantly different from the covered population — it may even result in biased population estimates. The situation becomes even more complex in the case of rare and elusive populations, which are in general difficult to survey (Sudman 1988). Besides the fact that they are not covered by standard sampling frames like business directories, they are usually hard to reach. Both the predominance of activity in the informal sector and the likelihood of fundamental differences between these and formal-sector firms necessitate regular, quality measurement of those units.

Virtually by definition, informal sector activity is not captured by standard sampling frames for business or enterprise surveys. Previous efforts to measure the informal sector have relied either on fully enumerated economic censuses or occur in conjunction with household surveys. Both methods are expensive and time-intensive, and while the latter method is frequently used, it does not capture informal sector units at their point of business, and so — while suited for measuring employment patterns — is not necessarily appropriate for measuring the population and activity of informal sector units as such.

One characteristic of populations such as informal sector units is that they are often distributed in clusters across a specific area, rather than being evenly spread across the area. Standard sampling designs, like stratified random sampling or cluster sampling would therefore require considerably large sample sizes as well as correspondingly large survey budgets. For all the above stated reasons we implement the current survey by using a stratified adaptive spatial cluster design (Thompson 1989, 1991).

The rest of the paper is organized as follows. Section 2 provides background on informal sector activity in Harare, Zimbabwe, which was the site of the survey pilot. Section 3 discusses the survey implementation. Section 3 discusses the calculation of design weights, while Section 4 lays out initial results. Section 5 concludes.

2. Background on the informal Sector in Harare Zimbabwe

Informal sector employment and activity is known to predominate in many developing economies. Zimbabwe is certainly no exception: The 2014 Labour Force and Child Labour Survey (LFCLS) conducted by Zimbabwe's national statistics authority (Zimstat) estimated that a full ninety-four percent of the working age population (older than fifteen years of age) held informal employment. A slight majority (52.5 percent) of these workers are female, and thus the sector also features a strong gender component. Over sixty percent of informal sector employment (when agriculture is excluded) is in the trade sectors, including both retail and wholesale: A great number of informal operators obtain a daily permit to sell goods or services legally.

The LFCLS used a stratified, two-stage sampling methodology, with enumeration area Primary Sampling Units (PSUs) with probabilities of selection proportional to the concentration of households from the 2012 population census. Second-stage sampling was conducted based on listings of households within the identified enumeration areas. In total, the study covered 419 enumeration areas and surveyed 10,475 households. The enumeration of households for second-stage selection took place in March 2014, while fieldwork took place in June of the same year.

The coverage and methodology used in the LFCLS are quite similar to those used in similar studies and methods laid out by the Delhi Group on Informal Sector Statistics (ILO 2013). These methodologies are commonly referred to as household-based methods. Methodologies that also visit enterprises based on household enumeration are known as combined household-enterprise methods.

Both such methods may not be suitable for the current aims of this survey. First, the population of interest for this survey is all informal sector units or enterprises operating in Harare. As such, it is conducted at the point of business and seeks to measure units' interaction with the business environment, their activity, and production. Each of these is of relevant policy interest and uniquely applicable at the unit level. Policymakers are often concerned with creating paths to formalization for informal sector activity, which may enhance overall productivity, and thus the unit of analysis as the informal business or activity is important.

Second, and relatedly, the definition of informality used by such labour force surveys is based on the type of activity and typically consists of those employees not receiving any benefits (nor covered under employment-related taxes), often including in-home work. Thus, informal work can take place in formal establishments, and presumably policy and research questions regarding these workers are different from those affecting informal sector enterprises per se.

Third, as informal sector activity is known to often be clustered, such two-stage sampling methods may be inefficient in capturing informal sector activity; moreover, they do not yield information on the location of these activities. Lastly, informal sector employment is also known to be transitory, in both location and time, and so an application of in-field second-stage sampling is a useful innovation, which is described in the following section on implementation.

3. Implementation

The first step was the construction of a spatial grid as our Primary Sampling Units (PSU) frame, as shown in Figure 1. The grid covered the total of municipal Harare, and each quadratic cell had a size of 200 by 200 meters. For the purposes of the study, Harare was divided into three strata of low, medium, and high concentration of informal sector activity based on expert assessment. The target population are the informal sector units in Harare, Zimbabwe, where informality is defined as the lack of registration with the Registrar of Companies. Area frames are well known from household surveys, though they commonly use administrative boundaries as the delimitation. However, the use of a regular grid as an area frame has a long tradition in Ecological surveys (Greig-Smith 1964), although its application to human populations is relatively rare.

To address the second problem, namely the clustered distribution of the target population across the survey area, we used adaptive cluster sampling. Adaptive stratified (cluster) sampling (Thompson 1990, 1991) is useful in the case of rare or elusive populations with a clustered distribution pattern. It is a sampling design in which the selection of some of the final sampling units depends on the value of a variable of interest observed during the survey. In adaptive sampling one selects usually a sample of starting units, and this sample then constitutes the start of each adaption process (usually called a network). If the number of units in a starting square exceeds a predefined threshold, all the other squares surrounding the starting square are surveyed. If the observed actual value does not exceed the threshold, only the starting cell is surveyed, and the network is called a network of size 1. If one of the surrounding squares exceed the threshold, then the squares surrounding this one are subsequently surveyed. This process is continued until either the network is exhausted, or an arbitrary cut-off point is defined. We defined this cut-off for the 4^{th} expansion, though it was never reached in fieldwork. One challenge with this approach is the strong degree of discretion left to the interviewer. This problem gets even more relevant if this design is implemented in a low-skill environment. However, the use of electronic data collection and monitoring tools enables the implementation of more challenging types of survey designs even in a low-skilled environment.

To make the survey design even more efficient, and to fully exploit the capabilities of the electronic data collection and survey management tools, we apply adaptive stratum allocation (Solomon and Zacks 1970). This strategy follows a two-phase design, with Phase 1 being the "learning" phase, which provides the data for the re-evaluation of the initial allocation, and a Phase 2, in which the allocation will be adjusted per the information generated in Phase 1.

The above described variable of interest, which defines the threshold for the expansion of the network will be the number of units encountered in the initially selected PSU. The final sample size, as well as the initial allocation was defined through a simulation. This simulation is implemented in R and uses the Shiny library. The latter allows for the development of a Graphical User Interface (GUI). The GUI is developed in a way such that it does not require any programming skills for future use, and can easily be applied by statistical agencies lacking the required skill base to develop an approach like this on their own.

As mentioned above, several pieces of information on informal sector activity are of interest for both policymakers and researchers, and thus a full-length questionnaire was developed, building on previous modules used by the World Bank to measure constraints to informal firms. This survey was fully implemented into the World Banks' Survey Solutions CAPI system and is referred to as the long-form survey. As the stratified adaptive cluster sampling methodology used requires full enumeration of all informal sector units within a PSU, units were randomly selected for the long-form survey using the CAPI system. This selection was conducted in real time using the CAPI system and thus addresses issues stemming from the transitory nature of many informal sector activities. All respondents that were not selected for the long-form were given a short-form questionnaire, which captured information on the type of activity, physical location, and the number of workers engaged; units that refused the long-form were given the option to take the short-form questionnaire. Outright refusals were also recorded, using enumerator observation of the activity and workers observed. Also in R and again with a Shiny GUI, we developed an enhanced Survey management interface, which allows for full control of the survey teams in the field. All the additional software components will fully integrate with Survey Solutions.

Implementation of the actual fieldwork can be daunting given the complicated nature of the sampling methodology. A series of training and piloting was conducted before the launch of the fieldwork. An initial training of enumerators and field management team took place in November

2016, followed by piloting. Based on feedback from this training and piloting, necessary changes were made to the questionnaire and CAPI script. A second round of training with the entire field team took place in March 2017, an intense full-day training seminar followed by piloting and de-briefing on the second day. A third and final (virtual) training was conducted in the first week of April 2017 to clear any outstanding issues and fine tune survey instrument and data collection methodology. Fieldwork was conducted between April to mid-June 2017. The fieldwork was implemented by Probe Market Research, a local survey contractor based in Harare.

During the data collection phase, a detailed monitoring protocol was put in place to ensure the integrity of the fieldwork methodology. In addition to supervision through assigned supervisors, every enumerator records his/her path using a tracking device installed on the CAPI tablet and submits to a centralized server. This tracking path is checked by overlaying it on mapping software to ensure that enumerators have fully covered the square assigned to them. This check is done daily, and for cases where the tracking path indicate below acceptable level of effort in listing informal firms, the enumerator is asked to re-survey the square.

4. Calculation of design weights

To estimate population parameters, weights are applied to survey samples. These weights are derived as the inverse of their probability of selection. For simple random sampling a weight is calculated as $w = \frac{1}{\pi} = \frac{N}{n}$, indicating that the corresponding sampling unit represents w number of population units. This is the standard approach applied to most of the surveys, though adjusted correspondingly to the type of design. One particular feature of this approach is that they can be calculated ahead of the survey, and are eventually only adjusted for non-response/underrepresentation. The corresponding Horwitz-Thompson (HT) estimator is then as follows.

$$\overline{y}_{pop} = \sum_{1}^{h} \overline{y}_h$$

with

$$\overline{y}_h = \sum_1^i \frac{y_{h,i}}{\pi_{i,h}}$$

Conversely in adaptive spatial cluster sampling weights are not known at the start. The reason for this is that the final sample size is not known precisely at the survey start, and neither is the network size.[1] Therefore, ACS requires weights to be calculated differently. In adaptive spatial cluster sampling one talks about networks. Networks can be of different sizes, denoted by m_i. The simplest network is the one with only a single node, namely the selected starting square. Probabilities for these networks are calculated as in a stratified simple random sample setting:

$$\pi_{i,h}^0 = \frac{n_h}{N_h}$$

with h indicating the corresponding stratum, n and N the corresponding sample and population size. However, things become more complicated with networks of size $m_i > 1$, which is the result if the initial square exceeds the threshold. In this case $\pi_{i,h}$ needs to account for the different selection probabilities at the different stages:

$$\pi_{i,h}^1 = 1 - \left[\frac{\binom{N_h - m_i}{n_i}}{\binom{N_h}{n_h}} \right]$$

The final probability $\pi_{i,h} = \pi_{i,h}^0 + \pi_{i,h}^1$ can then be applied in the familiar way to calculate population means, totals and standard errors. An additional adjustment needs to be made if a network crosses the stratum boundaries, and the starting squares are different, as well as when networks overlap. The latter did not apply to our case.

5. Results

One of the benefits of this sampling approach is that we can provide a statistically sound estimate of the number of informal establishments operating in Harare. This is important

[1] It is possible, to approximate the sample size and costs by setting up an empirical sampling simulation, as it was done in this project

information for policymakers in their effort to design policies targeted to informal sector firms. Per our estimate, there are roughly about 30,000 informal establishments in the city of Harare alone. To put this in context, the 2014 establishment census conducted by Zimstat shows that there are about 7,000 formally registered firms operating in Harare: This implies that the frequency of informal sector units is almost five times that of formal establishments.

The majority of informal establishments in Harare are own-account operated (55%) (Figure-2), run by just a single person, often the owner. Nevertheless, a small fraction (9%) engage 5 or more individuals. As an indication of substantial firm turnover in the sector, a significant share of firms is 5 years old or younger (77%), and about 36% have been in operation for a year or less. Interestingly, about 11% of these informal firms are mature, having been active for 11 or more years. It is not clear if these firms have been operating as an informal unit all along, or switched from formal to informal at some point in their operating life.

The re-selling of goods (i.e., retail trade) is the predominant activity of informal firms in Harare (figure-3), accounting for about 65% of our sample while about 16% of firms are engaged in manufacturing of some sort. This is in line with anecdotal observation that the sector is populated mostly by street traders and vendors. Consistent with this, about 47% of the establishment have no fixed location from which they operate; the remaining fraction of firms have fixed location of some form.

A clear majority (80%) informal firms in Harare have a single owner; about 18% report that they are owned by two individuals, with just 2 percent reporting more than two owners. As owners of informal firms are often the managers (and in most cases the only employee) of the company, owner characteristics is more important predictor of performance than is the case for larger and more formal firms. Unlike the common perception of informal activities as kinds running around the traffic lights selling chewing gums, majority of informal firms in Harare are owned by middle-aged individuals, with a median age of about 36 years (figure-4). In fact, for about 30%, the largest owner is 60 years old and above. About 95% of the largest owners of informal establishments in Harare have secondary school education or above, and about 25% have a university degree or some vocational education ((figure-5). Gender-wise, although information is missing for 20% of our sampling, the sector appears to be dominated mainly by male, with about 44% of firms reporting that the largest owner is male compared with 36% for female.

6. Concluding remark

In this pilot survey, we successfully implemented a novel approach to survey an important type of an elusive, nevertheless economically significant population, namely informal economic establishments. We have achieved this by an eclectic approach of combining statistically sound methodologies from other areas in statistics and new data collection technologies.

As the technology part was covered by free or open source technologies, as well as by the development of a user-friendly toolbox, which the authors are happy to share, we believe that this approach could also be implemented with statistical agencies in low skill environment on a regular and sustainable basis.

7. References

De Soto, Hernando. 1989. The Other Path: The Invisible Revolution in the Third World. New York: Harper and Row.

Greig-Smith, P. (1964) Quantitative plant ecology. (2nd ed.) Butterworths, London.

Harris, John, and Michael Todaro. 1970. "Migration, Unemployment and Development: A Two-Sector Analysis." American Economic Review 60, no. 1: 126–42.

International Labour Organisation (ILO). 2013. Measuring informality: a statistical manual on the informal sector and informal employment. International Labour Office. Geneva: ILO. ISBN 9789221273882; 9789221273899 (web pdf)

La Porta, R. and A. Shleifer. (2014). "Informality and Development". The Journal of Economic Perspectives. 28(3): 109-126.

Sudman, S., Sirken, M., Cowan, C. D. (1988). Sampling rare and elusive populations. Science240.4855: 991

Solomon, H., & Zacks, S. (1970). Optimal design of sampling from finite populations: A critical review and indication of new research areas. Journal of the American Statistical Association, 65(330), 653-677.

Thompson, S. K. (1990). Adaptive cluster sampling. Journal of the American Statistical Association, 85(412), 1050-1059.

Thompson, S. K. (1991). Stratified adaptive cluster sampling. Biometrika, 78(2), 389-397.

Zimbabwe National Statistics Agency (Zimstat). 2014. Labour Force Survey

Figure 1: PSU sampling frame

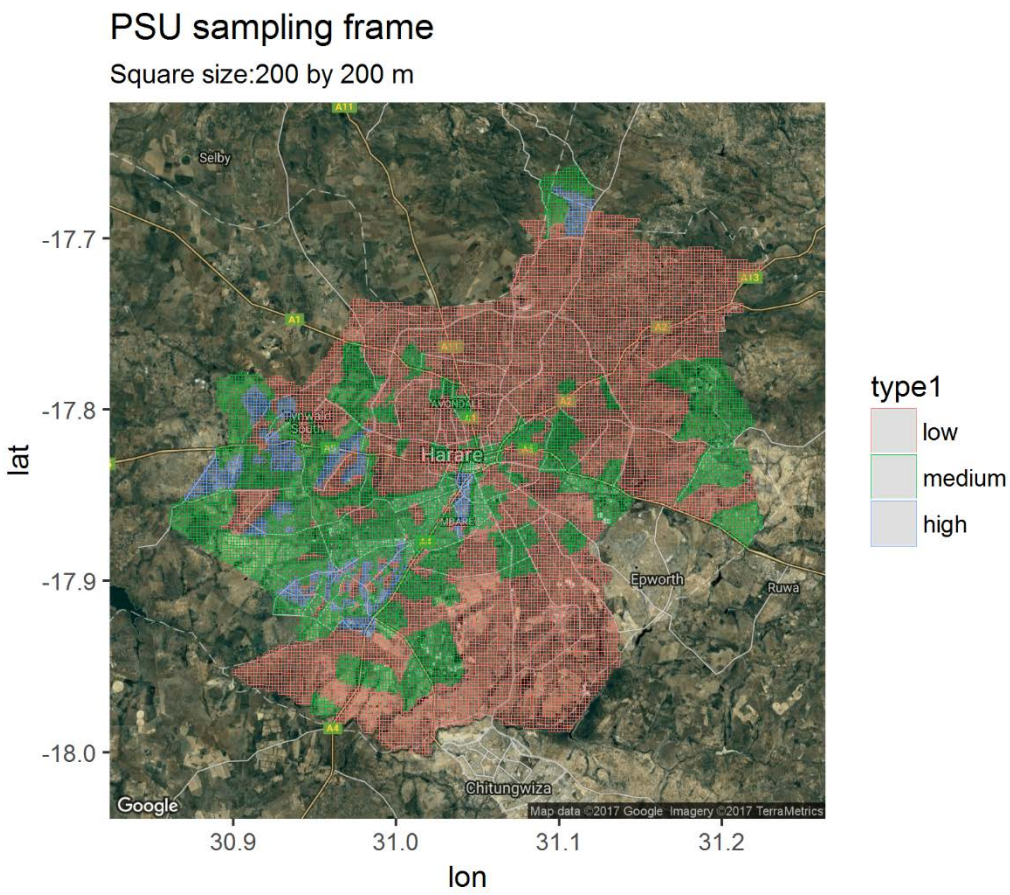

Figure 2: Age and Size of Informal Establishments in Harare

Figure 3: Sector and Location of Informal Establishments in Harare

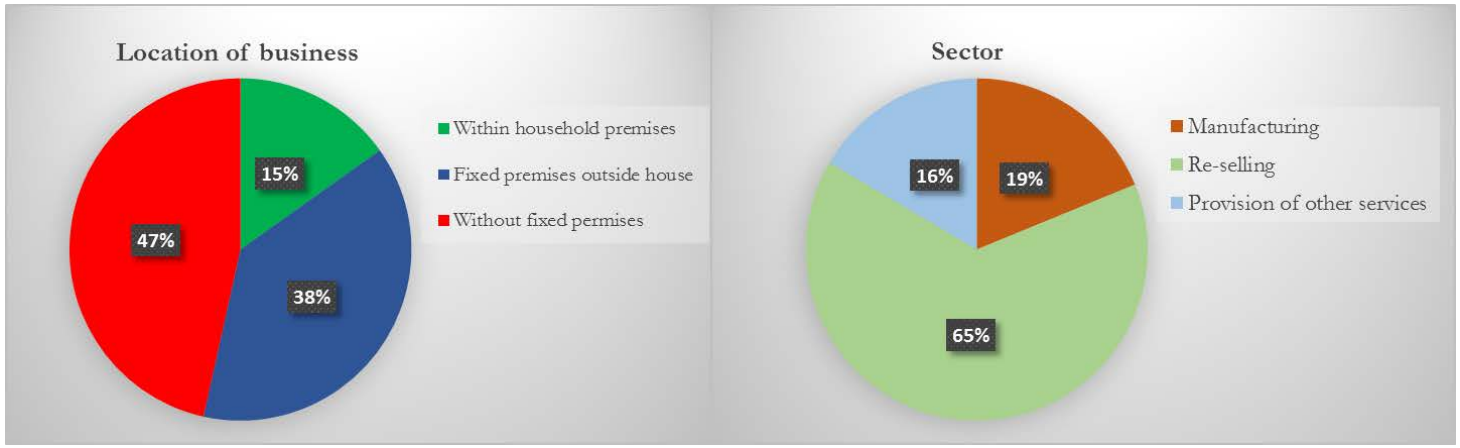

Figure 4: Age and Gender of the largest Owner

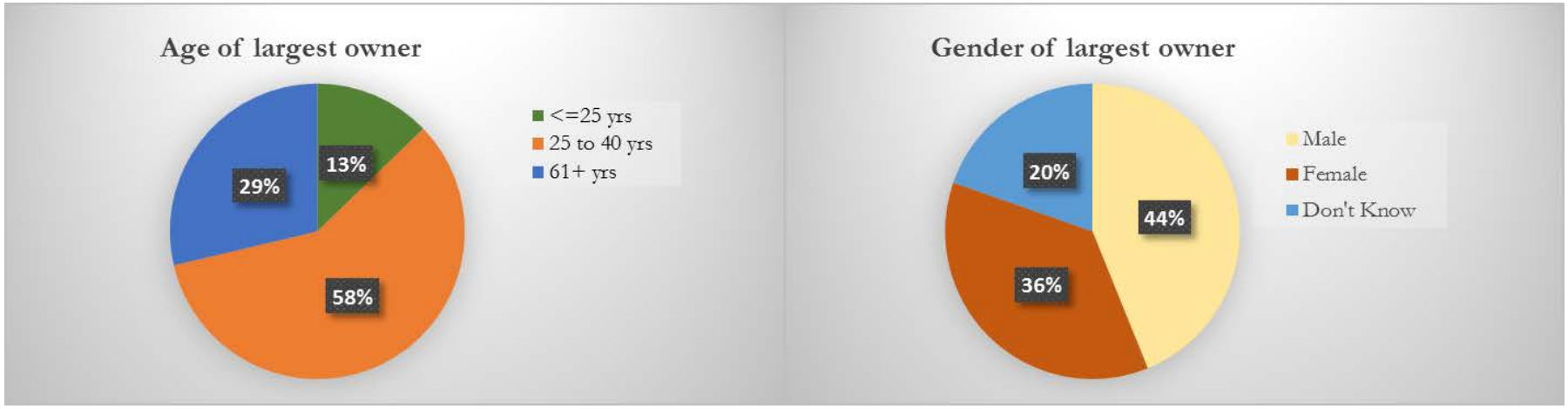

Figure 5: Education and alternative employment of the largest Owner